



Article

Deep Learning for Facial Beauty Prediction

Kerang Cao ¹ , Kwang-nam Choi ², Hoekyung Jung ³  and Lini Duan ^{1,*}

¹ Department of Computer Science and Engineering, Shenyang University of Chemical Technology, Shenyang 110000, China; caokerang@syuct.edu.cn

² NTIS Center, Korea Institute of Science and Technology Information, Seoul 34113, Korea; knchoi@kisti.re.kr

³ Department of Computer Engineering, Paichai University, Daejeon 35345, Korea; hkjung@pcu.ac.kr

* Correspondence: liniduan@163.com

Received: 14 July 2020; Accepted: 6 August 2020; Published: 10 August 2020



Abstract: Facial beauty prediction (FBP) is a burgeoning issue for attractiveness evaluation, which aims to make assessment consistent with human opinion. Since FBP is a regression problem, to handle this issue, there are data-driven methods for finding the relations between facial features and beauty assessment. Recently, deep learning methods have shown its amazing capacity for feature representation and analysis. Convolutional neural networks (CNNs) have shown tremendous performance on facial recognition and comprehension, which are proved as an effective method for facial feature exploration. Lately, there are well-designed networks with efficient structures investigated for better representation performance. However, these designs concentrate on the effective block but do not build an efficient information transmission pathway, which led to a sub-optimal capacity for feature representation. Furthermore, these works cannot find the inherent correlations of feature maps, which also limits the performance. In this paper, an elaborate network design for FBP issue is proposed for better performance. A residual-in-residual (RIR) structure is introduced to the network for passing the gradient flow deeper, and building a better pathway for information transmission. By applying the RIR structure, a deeper network can be established for better feature representation. Besides the RIR network design, an attention mechanism is introduced to exploit the inner correlations among features. We investigate a joint spatial-wise and channel-wise attention (SCA) block to distribute the importance among features, which finds a better representation for facial information. Experimental results show our proposed network can predict facial beauty closer to a human's assessment than state-of-the-arts.

Keywords: deep learning; facial beauty prediction; convolutional neural network

1. Introduction

As a burgeoning issue [1], facial beauty prediction (FBP) has attracted more and more attention from researchers and users, which is a comprehensive topic of face recognition [2,3] and comprehension [4–6]. An example of an FBP problem can be demonstrated in Figure 1. There are application potentials for FBP with attractiveness, such as makeup recommendation, and face beautification.

In FBP problem, facial features play an important role for assessment. After extraction, the features are explored and summarized for aggregate analysis. To find a better representation of facial features, there are various data-driven models for FBP with hand-crafted or adaptive learned descriptors [7–9]. With extracted features, these models perform the assessments with elaborate predictors, which are trained in a statistic manner.

Lately, deep learning has been proved as an efficient tool for signal and image processing [10–12]. The revival of deep learning methods, especially, convolutional neural networks (CNNs), provides a new perspective for FBP problem. CNN performs much better performances in

plentiful computer vision tasks than traditional methods [13–15], such as text localization [16], image classification [17], facial landmark regression and analysis [18,19], emotion recognition [20,21], time series forecasting [22], and semantic segmentation [23]. With the adaptive feature extraction and exploration, CNN demonstrates superior capacities on the high-level computer vision tasks. To our best knowledge, AlexNet [24] is the first CNN-based method for real-world image recognition, which was proposed for image classification task in ImageNet. After AlexNet, VGGNet [25] explored a deeper and wider design for better feature exploration and network performance. There are various effective network designs for better performance on different tasks, which lead to a blossom of deep learning.

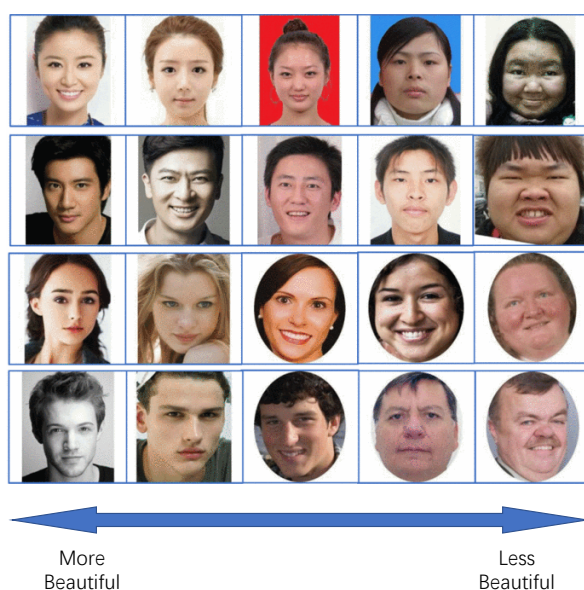


Figure 1. An example of facial beauty prediction. From top to the bottom, there are four kinds of facial images for prediction: Asian Female (AF), Asian Male (AM), Caucasian Female (CF), and Caucasian Male (CM). From left to the right, the predicted ranks are decreased gradually.

ResNet [26], proposed by He et al., has turned out to be a remarkable design pattern for CNN architectures. While building a deeper network, there is gradient vanishing problem which limits the performance. In ResNet, a shortcut is designed as bypath, connecting the inputs and outputs for better gradient transmission. The component design with shortcut is termed as residual block, which aims to learn the residual information from main path. Based on the residual learning, it is able to build a very deep network.

There are varieties of ResNet focusing on efficient network representation. In ResNeXt [27], group convolutions were introduced to introduce the cardinality of the network, which means the size of transformations. From the investigation, it is more efficient to improve the cardinality than depth and width. ResNeXt holds a similar structure to InceptionNet [28]. However, there are identical topology structures in ResNeXt, which reduce the design burden. With the splitting-transformation-merging strategy, ResNeXt achieved competitive performance compared to ResNet with much fewer parameters.

ResNeSt [29] is another effective design derived from ResNet. With channel separation and attention, ResNeSt block designs several identical cardinals in the computation unit. Similar to SENet [30] and SKNet [31], there are channel attentions in each cardinal unit for better feature map representation. Considering the inherent correlation of different features, ResNeSt has become state-of-the-art for image recognition and beyond.

Residual connection and its varieties demonstrate the superior performance of network representation ability. To build the network deeper and establish a more robust gradient flow, residual-in-residual (RIR) is proposed by grouping residual blocks with a higher level shortcut. With the high level residual connection, gradient and information will be transmitted from the

shallow layers to the deeper. RIR structure has proved its performance on image super-resolution [32], restoration [33], classification and other computer vision tasks, which turns out to be an efficient design pattern.

Besides effective network designs, some works focus on the inherent correlations of features. Channel attention in SENet is one of the successful mechanisms for finding better representation of features. In SENet, information from different channels is evaluated by global average pooling and is processed by several full connection layers. Besides channel attention, spatial attentions are introduced by considering the dual attention maps both on channel-wise and spatial-wise features. Non-local attention [34], which is a special pattern for global information consideration on features, has become a success for image restoration and comprehension.

This paper proposes a novel CNN design for the FBP problem. In the proposed network, we investigate an RIR block design with an attention mechanism. To build the network deeper, multi-level skip connections are introduced to compose a better gradient transmission flow. An attention mechanism is devised to find the inherent correlation among feature maps. In an attention mechanism, both channel-wise and spatial-wise attentions are considered for a better correlation representation. Experimental results show that our network holds a better performance than other CNN-based methods, which is more consistent with the assessment of humans.

The contributions of this paper can be demonstrated as follows:

- We propose a network for the facial beauty prediction (FBP) problem. Specifically, residual-in-residual (RIR) groups are designed for building a deeper network. To devise a better gradient transmission flow, multi-level skip connections are introduced.
- To find the inherent correlations among features, a joint spatial-wise and channel-wise attention mechanism is introduced for better feature comprehension.
- Experimental results demonstrate our network can achieve a better performance than other CNN-based methods and make the assessment more consistent with human opinion.

2. Related Works

2.1. Facial Beauty Prediction

Facial beauty acts as an essential influence factor in daily life. Facial beauty prediction (FBP) has attracted more and more attention from researchers, which is a composite study of psychology, computer science, evolutionary biology, and so forth. Recently, there are data-driven methods for the FBP issue, which adaptively extract the facial features and perform the analysis with defined mathematical models. The models are supposed to be consistent with a human's assessment as much as possible. Besides plentiful prediction methods, databases proposed for the FBP problem have also become a spotlight in this area. SCUT-FBP5500 [9], to our best knowledge, is one of the most popular open benchmarks widely used for evaluation. In SCUT-FBP5500, there are diverse face pictures containing both males and females. The ages of the persons vary from 15 to 60, which acquires a large interval for a richer representation of real facial beauty situation. There are 2000 Asian females, 2000 Asian males, 750 Caucasian males and 750 Caucasian females in SCUT-FBP5500, resized to 350×350 resolution. The scores are ranked by 60 volunteers between 1–5, meaning the attractiveness from low to high.

2.2. Convolutional Neural Networks

Convolutional neural networks (CNNs) are one of the most remarkable successes in the deep learning area, which demonstrate superior performances on a large amount of computer vision tasks, such as detection [35–37], denoising [38], recognition [39,40], and better feature representation [41–43]. To our best knowledge, AlexNet, as the champions of the ImageNet competition in 2012, is one of the successful CNN designs developed on GPU. After AlexNet, there have been numerous efficient designs for better performances, composed of wider or deeper networks with elaborate and fancy

layer connections. VGGNet, which is widely applied for various GAN-based works, is one of the representative design patterns for deep networks. InceptionNet proposed by Google is another effective design which achieved the first prize of ILSVRC 2014. In InceptionNet, the authors designed an inception module to improve the parameter utilization. The inception module applied 1×1 convolutional layers to organize the information across different channels. Features from convolutional layers with different filter sizes and max-pooling operations were aggregated by concatenation. By utilizing the Inception module, the network improved the performance and avoided the over-fitting with more branches. Furthermore, batch normalization (BN) and dropout strategies were applied in InceptionNet for training speed improvement.

Residual connection from ResNet is another well-known structure for network design. In VGGNet and InceptionNet, there is a limitation of network performance with the increase of network depth, which is caused by the gradient vanishing. To solve this issue, residual representation was introduced in ResNet. The blocks were designed based on the residual learning, which makes it easier to learn the mapping between inputs and outputs.

Densely connection [44], as another efficient design pattern for CNN, has become one of the popular choices for various tasks. Different from ResNet, DenseNet applied dense connections to connect features from shallow layers to deeper, which are more effective than residual connection. Furthermore, features are reused in DenseNet via the concatenation of channels, which saves the parameters with fewer computation costs.

3. Method

As shown in Figure 2, the proposed network holds a pyramid structure to progressively extract the feature from images. We devise a residual-in-residual group termed RIRG to build the network deeper. There are five stages in the network. For each stage, the feature maps are down-scaled by max-pooling operation as the output. Finally, a global average pooling is utilized to shrink the resolution to 1×1 . Let us denote the input image as \mathbf{I}^0 , then for i -th stage, the operations can be described as,

$$\begin{aligned} \mathbf{feat}^i &= RIRG_N^i(\dots RIRG_1^i(\mathbf{I}^i) \dots), \\ \mathbf{I}^{i+1} &= MaxPool(\mathbf{feat}^i), \end{aligned} \tag{1}$$

where $RIRG_n^i(\cdot)$ denotes the n -h RIRG in i -th stage, and $MaxPool(\cdot)$ denotes the max pooling operation.

The proposed network will be introduced in the following manner. Firstly, we will introduce the structure of the proposed RIRG. Moreover, the spatial-wise and channel-wise joint attention mechanism applied in RIRG will be demonstrated in detail, which is termed SCA. Finally, the settings of the proposed network will be described with discussions.

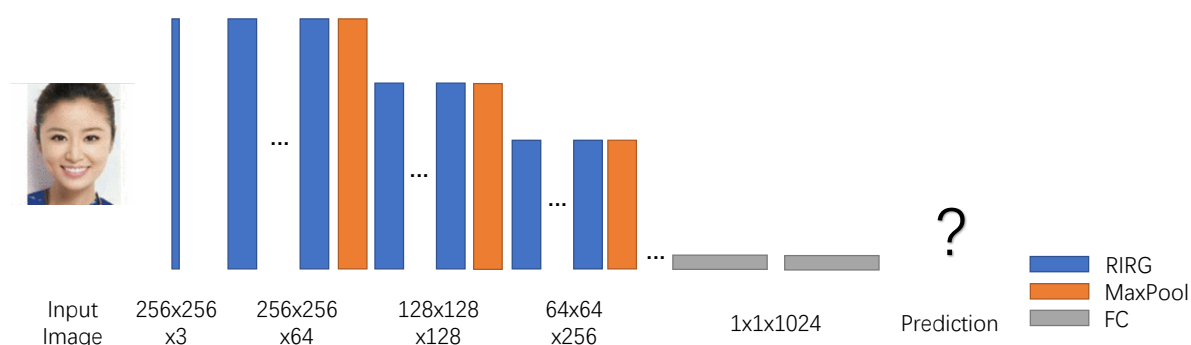


Figure 2. Network structure. There is a progressive design for facial feature exploration. With the increase of channel number, the resolution of feature maps will be decreased with a fixed ratio.

3.1. Residual-In-Residual Group

RIRGs are devised from the perspective that a deeper network will lead to better performance. Since residual connection can survive from the gradient vanishing problem, residual-in-residual connections are introduced to pass the shallow features and gradients to deeper with a long shortcut. There are multi-levels in RIRG to build the flow more effective. The design of RIRG is shown in Figure 3.

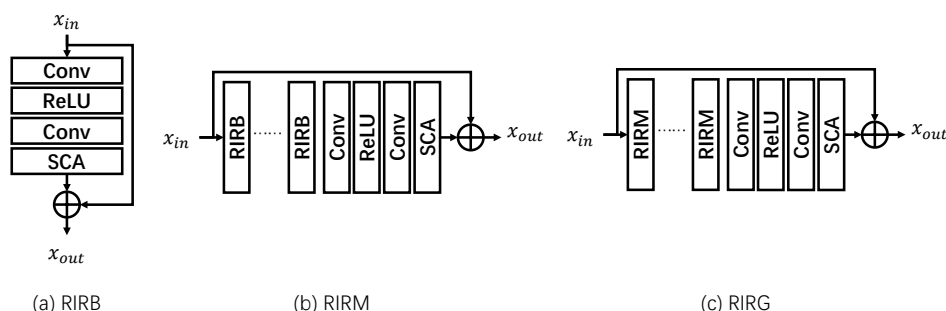


Figure 3. Structure of proposed RIRG. The RIRG is built in a residual-in-residual way for better information transmission.

The basic block in RIRG, termed RIRB, as shown in Figure 3a, is composed of two convolutional layers with a ReLU activation and SCA block. The residual connection in RIRB can be regarded as the first level skip path, which connects the features at a distance of two convolutional layers. By stacking RIRBs, the residual-in-residual module (RIRM) is designed with a padding structure. The second level skip path is introduced to RIRM connecting the input and output features. Beyond the second level skip connection, residual-in-residual group (RIRG) is devised in a recursive way like the RIRM, which stacks the RIRM as the main path with a same padding structure. The third level skip connection in RIRG crosses a large spacing of convolutional layers, which makes the shallow features become deeper more efficiently.

The RIRG holds a similar structure to dense connection. If we expand the three-level residual connections, and regard the stacked layers as an entire operation, then the features from different layers are densely connected via residual learning. On one hand, the densely-like design is able to deliver the features and information more efficiently. On the other hand, the mixture of dense and residual connections reuses the features with limited parameters and computation complexity, which indeed improves the network performance.

Although the RIR design can build an efficient information transmission pathway, it requires a large number of parameters and high computation complexity with the increase of network depth. From this point of view, we propose a modified convolution operation to substitute the vanilla layer. For each convolution step, there are two 1×1 convolution operations for channel squeeze and excitation, and one depth-wise convolution for spatial exploitation. The shrunken channel number is set as 32. With this substitution, the computation complexity and parameters will be substantially saved.

3.2. Spatial-Wise and Channel-Wise Attention

The proposed spatial- and channel-wise attention mechanism (SCA) is shown in Figure 4. As shown in the illustration, there are two dual paths finding the spatial-wise and channel-wise attentions separately. There is a convolutional layer to demonstrate the explore the correlation from features in general. After exploration, two parallel bypaths exploit the different attentions independently. From the channel-wise attention bypath, the information from different channels will be evaluated by global average pooling. After pooling, two full connection layers with a ReLU activation is introduced to dig out the inherent correlation. Finally, a Sigmoid activation is devised for the non-negativity.

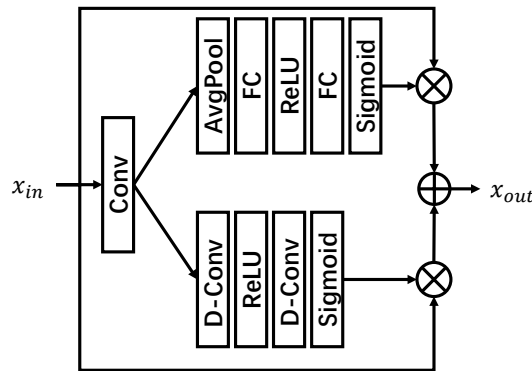


Figure 4. Structure of proposed spatial- and channel-wise attention mechanism (SCA). There are two paths for finding the spatial and channel attention jointly. After distribution, the addition of two features will be regarded as the output.

Similar to the channel-wise bypass, there are two convolutional layers with a ReLU activation to demonstrate the correlation of spatial-wise features, and a Sigmoid activation is applied at the end of the processing procedure. Different from the channel-wise bypass, there is no global pooling method for information evaluation.

After extraction, the two attentions are multiplied with the input features, and the addition is regarded as the final result. The operation of SCA can be described as,

$$\mathbf{x}^{SCA} = Conv(\mathbf{x}_{in}), \tag{2}$$

$$\mathbf{x}_C^{SCA} = \sigma(FC(ReLU(FC(AvgPool(\mathbf{x}^{SCA}))))), \tag{3}$$

$$\mathbf{x}_S^{SCA} = \sigma(DConv(ReLU(DConv(\mathbf{x}^{SCA}))), \tag{4}$$

$$\mathbf{x}_{out} = \mathbf{x}_C^{SCA} \otimes \mathbf{x}_{in} + \mathbf{x}_S^{SCA} \odot \mathbf{x}_{in}, \tag{5}$$

where $\sigma(\cdot)$ denotes the Sigmoid activation, \mathbf{x}_{in} , \mathbf{x}_{out} denote the input and output features separately. \otimes denotes the channel-wise multiplication, which allocates different weights to channels. \odot denotes the point-wise multiplication. $DConv(\cdot)$ denotes the depth-wise convolution. In the operation, \mathbf{x}_C^{SCA} is a tensor with size $1 \times 1 \times c$ and \mathbf{x}_S^{SCA} is a tensor with size $h \times w \times c$, where h , w , and c denotes the height, width and channel number of the size of \mathbf{x}^{SCA} .

3.3. Network Design

The entire network is designed as follows. Firstly, the input image is considered with size $256 \times 256 \times 3$. Then one convolutional layer expands the channel number to 64 and maintains the resolution. There are $N = 2$ RIRGs after the convolutional layers for feature exploration, and a max-pooling operation is applied to decrease the feature size. For each RIRG, there are five RIRMs, and five RIRB for each RIRM. There are $K = 2$ stages in the network. In each stage, the channel number will be expanded by one convolutional layer at the beginning, and the size of features is halved by max-pooling at the end. After the stages, there is a global average pooling step to resize the tensors as $1 \times 1 \times 1024$. Two fully connection (FC) layers with a ReLU activation is introduced to perform the prediction, and the output size of final FC is one, demonstrating the rank of facial beauty prediction.

4. Experiment

The network is trained on the SCUT-FBP5500 dataset. To our best knowledge, it is the largest dataset for FBP problem up to now. We train the network for 1000 iteration with batch size as $b = 25$. The parameters are updated by Adam optimizer with learning rate $lr = 1e - 4$, which is halved for every 200 iterations. We choose L1-loss as the loss function. Notice that the input size of the

SCUT-FBP5500 dataset is 224×224 , we rescale the image size to 256×256 by bicubic interpolation for training and testing.

4.1. Results

We conduct the comparison with diverse methods including geometric feature based and deep learning based methods—Linear Regression, Gaussian Regression, SVR, AlexNet, ResNet-18, and ResNeXt-50. The measurement indexes are chosen as Pearson Correlation, MAE, and RMSE. The dataset is split with the ratio 0.6 for training and 0.4 for testing. That is, 60% instances of the dataset are randomly chosen for training and the other 40% are for testing. The results are shown in Table 1.

PC demonstrates the Pearson Correlation, which is the higher the better. From the table, our network performs better than other works. With the higher PC, our proposed network is more consistent with human opinion, which shows the effectiveness of the proposed structure design.

Table 1. Performance comparison of different methods by 60–40% splitting.

| Methods | PC \uparrow | MAE \downarrow | RMSE \downarrow | Square Deviation \downarrow |
|-----------------|---------------|-------------------------------------|-------------------|-------------------------------|
| LR [9] | 0.5948 | 0.4289 \pm 1.50 | 0.5531 | 2.25 |
| GR [9] | 0.6738 | 0.3914 \pm 1.42 | 0.5085 | 2.03 |
| SVR [9] | 0.6668 | 0.3898 \pm 1.41 | 0.5152 | 1.99 |
| AlexNet [24] | 0.8298 | 0.2938 \pm 1.23 | 0.3819 | 1.53 |
| ResNet-18 [26] | 0.8513 | 0.2818 \pm 1.21 | 0.3703 | 1.48 |
| ResNeXt-50 [27] | 0.8777 | 0.2518 \pm 1.20 | 0.3325 | 1.45 |
| Ours | 0.8780 | 0.2517 \pm 0.65 | 0.3320 | 0.43 |

To further testify the network capacity, we perform the comparison via 5-fold cross validation, which holds 80–20% splitting for each fold. The results and average for each fold are shown in Table 2. From the table, our performance is better than state-of-the-arts.

Table 2. Performance comparison of the cross validation.

| PC \uparrow | 1 | 2 | 3 | 4 | 5 | Avg |
|-------------------------------|---------------|---------------|---------------|---------------|---------------|---------------|
| AlexNet [24] | 0.8667 | 0.8645 | 0.8615 | 0.8678 | 0.8566 | 0.8634 |
| R ³ CNN [45] | 0.8873 | 0.8741 | 0.8856 | 0.8906 | 0.8779 | 0.8831 |
| ResNet [26] | 0.8847 | 0.8792 | 0.8929 | 0.8932 | 0.9004 | 0.8900 |
| ResNeXt [27] | 0.8985 | 0.8932 | 0.9016 | 0.8990 | 0.9064 | 0.8997 |
| Ours | 0.8990 | 0.8939 | 0.9020 | 0.8999 | 0.9067 | 0.9003 |
| MAE \downarrow | 1 | 2 | 3 | 4 | 5 | Avg |
| AlexNet [24] | 0.2633 | 0.2605 | 0.2681 | 0.2609 | 0.2728 | 0.2651 |
| R ³ CNN [45] | 0.2436 | 0.2456 | 0.2428 | 0.2409 | 0.2451 | 0.2436 |
| ResNet [26] | 0.2480 | 0.2459 | 0.2430 | 0.2383 | 0.2383 | 0.2419 |
| ResNeXt [27] | 0.2306 | 0.2285 | 0.2260 | 0.2349 | 0.2258 | 0.2291 |
| Ours | 0.2300 | 0.2284 | 0.2257 | 0.2345 | 0.2251 | 0.2287 |
| RMSE \downarrow | 1 | 2 | 3 | 4 | 5 | Avg |
| AlexNet [24] | 0.3408 | 0.3449 | 0.3583 | 0.3438 | 0.3576 | 0.3481 |
| R ³ CNN [45] | 0.3155 | 0.3328 | 0.3227 | 0.3140 | 0.3294 | 0.3229 |
| ResNet [26] | 0.3258 | 0.3286 | 0.3184 | 0.3107 | 0.2994 | 0.3166 |
| ResNeXt [27] | 0.3025 | 0.3084 | 0.3016 | 0.3044 | 0.2918 | 0.3017 |
| Ours | 0.3020 | 0.3081 | 0.3013 | 0.3039 | 0.2916 | 0.3014 |
| Square Deviation \downarrow | 1 | 2 | 3 | 4 | 5 | Avg |
| AlexNet [24] | 1.5203 | 1.5255 | 1.5304 | 1.5225 | 1.5809 | 1.5359 |
| ResNet [26] | 1.4703 | 1.4731 | 1.4809 | 1.4720 | 1.4850 | 1.4762 |
| ResNeXt [27] | 1.4495 | 1.4506 | 1.4552 | 1.4533 | 1.4580 | 1.4533 |
| Ours | 1.4308 | 1.4350 | 1.4401 | 1.4420 | 1.4500 | 1.4395 |

Furthermore, we analyze the distribution of predicted scores from our network, which is shown in Figure 5. The yellow points are frequencies of prediction, while the blue points denotes the ground-truth. From the visualization illustrations, the score of male and female are in accordance with the normal distribution. There is a shift on mean value of male and female predictions. We hold the notion that the shift is from the bias of sexuality. To prove the hypothesis that our prediction accords with the normal distribution, we use the Anderson-Darling test for evaluation. It is a modification of the Kolmogorov-Smirnov (K-S) test and gives more weight to the tails than does the K-S test. We make the hypothesis H_0 : samples follow the normal distribution; and the H_1 : samples do not follow the normal distribution. After the A-D test, we get the significance level $\alpha = 0.043$ and the critical value $C_{A-D} = 0.75$. Since $\alpha \leq C_{A-D}$, we cannot reject the hypothesis H_0 . From this point of view, we hold the notion that the prediction values follow the normal distribution.

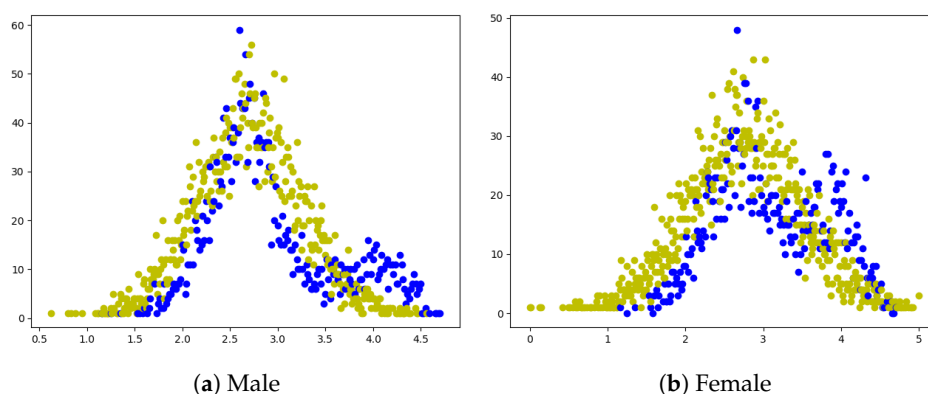


Figure 5. Visualization of prediction distribution. The blue points are the ranks of ground-truth. The yellow points are the ranks of prediction. The X-axis denotes the rank, and Y-axis is the frequency.

4.2. Ablation Study

Investigation on Network Design. To show the performance of network design, we make the experiments on different settings of block numbers. Specifically, K_m, K_g denote the block and module number in RIRM and RIRG separately. The results are shown in Table 3. From the table, the depth of network is one of the most important factors of network prediction performance. When the K_m and K_g are lower, the PC, MAE and RMSE will be worse than the longer network. This accords with the intuition that a wider and deeper network has better representation ability and processes features more effectively. Specifically, K_m and K_g have a similar influence on the performance. Singly adjusting the K_m and K_g has a similar effect on the results.

Table 3. Ablation on network design.

| Method | PC↑ | MAE↓ | RMSE↓ |
|--------------------|---------------|---------------|---------------|
| $K_m = 5, K_g = 5$ | 0.8780 | 0.2517 | 0.3320 |
| $K_m = 3, K_g = 5$ | 0.8491 | 0.2905 | 0.3755 |
| $K_m = 5, K_g = 3$ | 0.8480 | 0.2892 | 0.3749 |
| $K_m = 3, K_g = 3$ | 0.8301 | 0.2929 | 0.3800 |

Investigation on Attention Mechanism. To show the performance of the proposed SCA, we conduct the comparison with and without SCA blocks. The results are shown in Table 4. From the table, the attention mechanism leads to a shallow performance improvement due to the restricted parameters. SCA considers the channel attention and spatial attention jointly, which finds the correlations from two perspectives for better consideration.

Table 4. Ablation on attention mechanism.

| SCA | PC \uparrow | MAE \downarrow | RMSE \downarrow |
|-----|---------------|------------------|-------------------|
| w | 0.8780 | 0.2517 | 0.3320 |
| w/o | 0.8778 | 0.2525 | 0.3322 |

4.3. Discussion

Comparison on ResNeXt-50. In this paper, our network achieves a better performance than ResNeXt-50. Our network holds 6.75 M parameters and 34.25 GFlops. ResNeXt-50 has 25.03 M parameters and 5.56 GFlops. From the comparison, our network is lighter than ResNeXt-50. Although our network is much more deeper than other works, the well-designed convolution operation can prevent the plentiful number of parameters and computation complexity.

Effectiveness of SCA. In this paper, we propose an attention mechanism termed as SCA. Since there is only one convolution layer and some depth-wise convolutions in SCA, it is a simple but effective design for finding the inherent correlation of feature maps. There are few parameters in SCA with lower computation complexity. From this point of view, it can give a performance boost with a little increase on complexity. From this perspective, SCA is an efficient component for performance boost.

Comparison on Effective Network Architecture Designs. Recently, there are different effective network architecture designs for feature exploitation, such as ResNet and several extensions, DenseNet, MobileNet series and SqueezeNet. These works concentrate on different block designs for effective feature exploitation. However, the choreographed works are concentrating on building a deeper or wider structure for better performance, which lack to build a more efficient information transmission pathway. The main difference of different networks is the inside blocks, and almost all the networks are modified based on ResNet-50 or ResNet-101, which provide a fixed information transmission pathway for fair performance comparison. To address this issue, we introduce the RIR structure for better information transmission. Furthermore, these lightweight works focus on different blocks, but do not consider the correlations of features. In this paper, the attention mechanism is introduced to find the inherent correlations for better feature representation, which is termed as SCA.

Threats to Validity. In this paper, we propose a novel network for the FBP problem. However, the improved performance is limited by the number of parameters. A deeper or wider network will lead to better performance, while it will also produce a high computation complexity. Considering the threats to internal validity, the vital important element is the network depth. From the ablation study, with the increase of network depth, the performance will be improved at the same time. There are two aspects about the threats to external validity. On one hand, the labels for training are assessed by some students in a specific society, which may cannot cover a common opinion. On the other hand, the trained images are selected from Asian and Caucasian people, which may lead to a bias on the diversity.

5. Conclusions

In this paper, we proposed a novel network for the facial beauty prediction problem. Traditional networks focus on the effective block designs with a deeper or wider network for better performance, which almost neglect the efficient information transmission pathway and the correlations of features. To address these issues, we proposed a three-level residual-in-residual structure, termed RIRG, for better information transmission. Since RIRG was designed in a recursive way for multi-level residual connections, it could provide a more efficient information and gradient transmission style. Furthermore, a joint spatial and channel attention mechanism—SCA—was introduced in this paper for finding the inherent correlations of features, which is a tiny component with few parameters for performance improvement. The experimental results showed that our proposed network achieved a better performance than other works with restricted parameters. Further, we will find more datasets

with higher diversity, and compare our works with more recent works. Meanwhile, we will also tend to build a novel dataset for the FBP problem with more cultures, mentalities, traditions and economic status.

Author Contributions: Methodology, K.C.; Software, K.-n.C.; Writing—original draft, H.J.; Writing—review & editing, L.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Science Foundation of Shenyang University of Chemical Technology under grant No. LQ2020020.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gan, J.; Jiang, K.; Tan, H.; He, G. Facial Beauty Prediction Based on Lighted Deep Convolution Neural Network with Feature Extraction Strengthened. *Chin. J. Electron.* **2020**, *29*, 312–321. [[CrossRef](#)]
2. Gan, J.; Zhai, Y.; Wang, B. Unconstrained Facial Beauty Prediction Based on Multi-scale K-Means. *Chin. J. Electron.* **2017**, *26*, 548–556. [[CrossRef](#)]
3. Gunes, H.; Piccardi, M.; Jan, T. Comparative beauty classification for pre-surgery planning. In Proceedings of the 2004 IEEE International Conference on Systems, Man and Cybernetics, The Hague, The Netherlands, 10–13 October 2004; Volume 3, pp. 2168–2174.
4. Wassermann, S.; Seufert, M.; Casas, P.; Gang, L.; Li, K. Let me Decrypt your Beauty: Real-time Prediction of Video Resolution and Bitrate for Encrypted Video Streaming. In Proceedings of the 2019 Network Traffic Measurement and Analysis Conference (TMA), Paris, France, 19–21 June 2019; pp. 199–200.
5. Workman, S.; Souvenir, R.; Jacobs, N. Understanding and Mapping Natural Beauty. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5590–5599.
6. Chen, F.; Xiao, X.; Zhang, D. Data-Driven Facial Beauty Analysis: Prediction, Retrieval and Manipulation. *IEEE Trans. Affect. Comput.* **2018**, *9*, 205–216. [[CrossRef](#)]
7. Shi, S.; Gao, F.; Meng, X.; Xu, X.; Zhu, J. Improving Facial Attractiveness Prediction via Co-attention Learning. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 4045–4049.
8. Gan, J.; Xiang, L.; Zhai, Y.; Mai, C.; He, G.; Zeng, J.; Bai, Z.; Donida Labati, R.; Piuri, V.; Scotti, F. 2M BeautyNet: Facial Beauty Prediction Based on Multi-Task Transfer Learning. *IEEE Access* **2020**, *8*, 20245–20256. [[CrossRef](#)]
9. Liang, L.; Lin, L.; Jin, L.; Xie, D.; Li, M. SCUT-FBP5500: A Diverse Benchmark Dataset for Multi-Paradigm Facial Beauty Prediction. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 1598–1603.
10. Yu, Z.; Raschka, S. Looking Back to Lower-Level Information in Few-Shot Learning. *Information* **2020**, *11*, 345. [[CrossRef](#)]
11. Gibson, J.; Oh, H. Mutual Information Loss in Pyramidal Image Processing. *Information* **2020**, *11*, 322. [[CrossRef](#)]
12. Susilo, B.; Sari, R. Intrusion Detection in IoT Networks Using Deep Learning Algorithm. *Information* **2020**, *11*, 279. [[CrossRef](#)]
13. Zhang, Y.; Ding, W.; Liu, C. Summary of Convolutional Neural Network Compression Technology. In Proceedings of the 2019 IEEE International Conference on Unmanned Systems (ICUS), Beijing, China, 17–19 October 2019; pp. 480–483.
14. Devi, N.; Borah, B. Cascaded pooling for Convolutional Neural Networks. In Proceedings of the 2018 Fourteenth International Conference on Information Processing (ICINPRO), Bangalore, India, 21–23 December 2018; pp. 1–5.
15. Wang, Y.; Li, Y.; Song, Y.; Rong, X. Facial Expression Recognition Based on Random Forest and Convolutional Neural Network. *Information* **2019**, *10*, 375. [[CrossRef](#)]
16. Almakky, I.; Palade, V.; Ruiz-Garcia, A. Deep Convolutional Neural Networks for Text Localisation in Figures From Biomedical Literature. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; pp. 1–5.

17. Wan, S.; Gong, C.; Zhong, P.; Pan, S.; Li, G.; Yang, J. Hyperspectral Image Classification With Context-Aware Dynamic Graph Convolutional Network. *IEEE Trans. Geosci. Remote Sens.* **2020**. [[CrossRef](#)]
18. Lou, G.; Shi, H. Face image recognition based on convolutional neural network. *China Commun.* **2020**, *17*, 117–124. [[CrossRef](#)]
19. Stephen, O.; Maduh, U.J.; Ibrokchimov, S.; Hui, K.L.; Abdulhakim Al-Absi, A.; Sain, M. A Multiple-Loss Dual-Output Convolutional Neural Network for Fashion Class Classification. In Proceedings of the 2019 21st International Conference on Advanced Communication Technology (ICACT), PyeongChang, Korea, 17–20 February 2019; pp. 408–412.
20. Yu, D.; Sun, S. A Systematic Exploration of Deep Neural Networks for EDA-Based Emotion Recognition. *Information* **2020**, *11*, 212. [[CrossRef](#)]
21. Xiang, Z.; Dong, X.; Li, Y.; Yu, F.; Xu, X.; Wu, H. Bimodal Emotion Recognition Model for Minnan Songs. *Information* **2020**, *11*, 145. [[CrossRef](#)]
22. Cheng, Y.; Liu, Z.; Morimoto, Y. Attention-Based SeriesNet: An Attention-Based Hybrid Neural Network Model for Conditional Time Series Forecasting. *Information* **2020**, *11*, 305. [[CrossRef](#)]
23. Nascimento, J.C.; Carneiro, G. One shot segmentation: Unifying rigid detection and non-rigid segmentation using elastic regularization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**. [[CrossRef](#)] [[PubMed](#)]
24. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, Nevada, 3–6 December 2012; pp. 1097–1105.
25. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
27. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5987–5995.
28. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv* **2016**, arXiv:1602.07261.
29. Zhang, H.; Wu, C.; Zhang, Z.; Zhu, Y.; Zhang, Z.; Lin, H.; Sun, Y.; He, T.; Mueller, J.; Manmatha, R.; et al. ResNeSt: Split-Attention Networks. *arXiv* **2020**, arXiv:2004.08955.
30. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
31. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective Kernel Networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2019; pp. 510–519.
32. Wang, H.; Su, D.; Liu, C.; Jin, L.; Sun, X.; Peng, X. Deformable Non-Local Network for Video Super-Resolution. *IEEE Access* **2019**, *7*, 177734–177744. [[CrossRef](#)]
33. Lin, K.; Jia, C.; Zhao, Z.; Wang, L.; Wang, S.; Ma, S.; Gao, W. Residual in Residual Based Convolutional Neural Network In-loop Filter for AVS3. In Proceedings of the 2019 Picture Coding Symposium (PCS), Ningbo, China, 12–15 November 2019; pp. 1–5.
34. Xie, W.; Zhang, J.; Lu, Z.; Cao, M.; Zhao, Y. Non-Local Nested Residual Attention Network for Stereo Image Super-Resolution. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 2643–2647.
35. Okaishi, W.A.; Zaarzne, A.; Slimani, I.; Atouf, I.; Benrabh, M. A Traffic Surveillance System in Real-Time to Detect and Classify Vehicles by Using Convolutional Neural Network. In Proceedings of the 2019 International Conference on Systems of Collaboration Big Data, Internet of Things Security (SysCoBioTS), Casablanca, Morocco, 12–13 December 2019; pp. 1–5.
36. Pak, J.; Kim, M. Convolutional Neural Network Approach for Aircraft Noise Detection. In Proceedings of the 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC), Okinawa, Japan, 11–13 February 2019; pp. 430–434.

37. Lu, S.; Feng, J.; Wu, J. A Time Weight Convolutional Neural Network for Positioning Internal Detector. In Proceedings of the 2019 Chinese Control And Decision Conference (CCDC), Nanchang, China, 3–5 June 2019; pp. 4666–4669.
38. Sun, G.; Wang, Y.; Sun, C. Fault Diagnosis of Planetary Gearbox Based on Signal Denoising and Convolutional Neural Network. In Proceedings of the 2019 Prognostics and System Health Management Conference (PHM-Paris), Paris, France, 2–5 May 2019; pp. 96–99.
39. Lee, K.; Chae, S.; Park, H. Optimal Time-Window Derivation for Human-Activity Recognition Based on Convolutional Neural Networks of Repeated Rehabilitation Motions. In Proceedings of the 2019 IEEE 16th International Conference on Rehabilitation Robotics (ICORR), Toronto, ON, Canada, 24–28 June 2019; pp. 583–586.
40. Guha, S.R.; M. Rafizul Haque, S. Convolutional Neural Network Based Skin Lesion Analysis for Classifying Melanoma. In Proceedings of the 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh, 24–25 December 2019; pp. 1–5.
41. Alani, A. Arabic Handwritten Digit Recognition Based on Restricted Boltzmann Machine and Convolutional Neural Networks. *Information* **2017**, *8*, 142. [[CrossRef](#)]
42. Yu, Y.; Lin, H.; Meng, J.; Wei, X.; Zhao, Z. Assembling Deep Neural Networks for Medical Compound Figure Detection. *Information* **2017**, *8*, 91. [[CrossRef](#)]
43. Peng, M.; Wang, C.; Chen, T.; Liu, G. NIRFaceNet: A Convolutional Neural Network for Near-Infrared Face Identification. *Information* **2016**, *7*, 61. [[CrossRef](#)]
44. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
45. Lin, L.; Liang, L.; Jin, L. Regression Guided by Relative Ranking Using Convolutional Neural Network (R3CNN) for Facial Beauty Prediction. *IEEE Trans. Affect. Comput.* **2019**, *1*. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).